Review TRENDS in Ecology and Evolution Vol.19 No.2 February 2004

Model selection in ecology and evolution

Jerald B. Johnson[1] and Kristian S. Omland[2]

[1]Conservation Biology Division, National Marine Fisheries Service, 2725 Montlake Boulevard East, Seattle, WA 98112, USA
[2]Vermont Cooperative Fish & Wildlife Research Unit, School of Natural Resources, University of Vermont, Burlington, VT 05405, USA

Recently, researchers in several areas of ecology and evolution have begun to change the way in which they analyze data and make biological inferences. Rather than the traditional null hypothesis testing approach, they have adopted an approach called model selection, in which several competing hypotheses are simultaneously confronted with data. Model selection can be used to identify a single best model, thus lending support to one particular hypothesis, or it can be used to make inferences based on weighted support from a complete set of competing models. Model selection is widely accepted and well developed in certain fields, most notably in molecular systematics and mark–recapture analysis. However, it is now gaining support in several other areas, from molecular evolution to landscape ecology. Here, we outline the steps of model selection and highlight several ways that it is now being implemented. By adopting this approach, researchers in ecology and evolution will find a valuable alternative to traditional null hypothesis testing, especially when more than one hypothesis is plausible.

Science is a process for learning about nature in which competing ideas about how the world works are evaluated against observations [1]. These ideas are usually expressed first as verbal hypotheses, and then as mathematical equations, or models. Models depict biological processes in simplified and general ways that provide insight into factors that are responsible for observed patterns. Hence, the degree to which observed data support a model also reflects the relative support for the associated hypothesis.

Two basic approaches have been used to draw biological inferences. The dominant paradigm is to generate a null hypothesis (typically one with little biological meaning [2]) and ask whether the hypothesis can be rejected in light of observed data. Rejection occurs when a test statistic generated from observed data falls beyond an arbitrary probability threshold (usually $P < 0.05$), which is interpreted as tacit support for a biologically more meaningful alternative hypothesis. Hence, the actual hypothesis of interest (the alternative hypothesis) is accepted only in the sense that the null hypothesis is rejected.

By contrast, model selection offers a way to draw inferences from a set of multiple competing hypotheses. Model selection is grounded in likelihood theory, a robust framework that supports most modern statistical approaches. Moreover, this approach is rapidly gaining support across several fields in ecology and evolution as a preferred alternative to null hypothesis testing [1,3,4]. Advocates of model selection argue that it has three primary advantages. First, practitioners are not restricted to evaluating a single model where significance is measured against some arbitrary probability threshold. Instead, competing models are compared to one another by evaluating the relative support in the observed data for each model. Second, models can be ranked and weighted, thereby providing a quantitative measure of relative support for each competing hypothesis. Third, in cases where models have similar levels of support from the data, model averaging can be used to make robust parameter estimates and predictions. Here, we review the steps of model selection, overview several fields where model selection is commonly used, indicate how model selection could be more broadly implemented and, finally, discuss caveats and areas of future development in model selection (Box 1).

How model selection works

Generating biological hypotheses as candidate models

Model selection is underpinned by a philosophical view that understanding can best be approached by simultaneously weighing evidence for multiple working hypotheses [1,3,5]. Consequently, the first step in model selection lies in articulating a reasonable set of competing hypotheses. Ideally, this set is chosen before data collection and represents the best understanding of factors thought to be involved in the process of interest. Hypotheses that originate in verbal or graphical form must be translated to mathematical equations (i.e. models) before being fit to

Box 1. The big picture

- Biologists rely on statistical approaches to draw inferences about biological processes.
- In many fields, the approach of null hypothesis testing is being replaced by model selection as a means of making inferences.
- Under the model selection approach, several models, each representing one hypothesis, are simultaneously evaluated in terms of support from observed data.
- Models can be ranked and assigned weights, providing a quantitative measure of relative support for each hypothesis.
- Where models have similar levels of support, model averaging can be used to make robust parameter estimates and predictions.

Corresponding author: Jerald B. Johnson (jerry.johnson@noaa.gov).

Box 2. From multiple working hypotheses to a set of candidate models

To use model selection, verbal hypotheses must be translated to mathematical models. Ideally, the parameters of such models have direct biological interpretation, but translating hypotheses to meaningful models (as opposed to statistically arbitrary models, e.g. ANOVA or linear regression) is not always intuitive. Hence, we offer some guidance about how to get from multiple working hypotheses to a set of candidate models [2,6].

The first step is to specify variables in the model. Variables should correspond directly to causal factors outlined in the verbal hypotheses. The second step is to decide on the functions that define the relationship between independent variables and the response variable in terms of mathematical operators and parameters. In fields where model selection is commonly used (Box 5), appropriate functions can be found in published literature or tailored software [45,46]. In other fields, suitable models can be found in theoretical literature or borrowed from other disciplines. The third step is to define the error structure of the model.

Generating hypotheses and translating them to models is an iterative process. For example, one hypothesis might seem to be equally well depicted by two or more models, including different error structures. In such cases, the verbal rendition of the hypothesis must be refined so that there is a one-to-one mapping from hypothesis to model. This can lead to an increase in the number of working hypotheses; however, care should be taken not to include models with functional relationships among variables that are not interpretable. In this regard, model selection differs from data dredging, where the analyst explores all possible models regardless of the interpretability of their functions, or continues to develop models to be tested after analysis is underway [3].

Ultimately, the number of candidate models should be small (some argue, on philosophical grounds, that this should be fewer than 20 [3]). The guiding principle at this step is to avoid generating so many models that spurious findings become likely. Moreover, one should avoid relying on computing power to fit all available models in lieu of identifying a *bona fide* candidate set.

data [1,6]. Translating hypotheses to models requires identifying variables and selecting mathematical functions that depict the biological processes through which those variables are related (Box 2).

Fitting models to data

Once a set of candidate models is specified, each model must be fit to the observed data. At an early stage of the analysis, one can examine the goodness-of-fit of the most heavily parameterized (i.e. global) model in the candidate set [3]. Such goodness-of-fit can be assessed using conventional statistical tests (e.g. χ^2 tests or G-tests) [7] or a PARAMETRIC BOOTSTRAP procedure (see Glossary). If the global model provides a reasonable fit to the data, then the analysis proceeds by fitting each of the models in the candidate set to the observed data using the method of MAXIMUM LIKELIHOOD or the method of LEAST SQUARES.

Selecting a best model or best set of models

Model selection is frequently employed as a way to identify the model that is best supported by the data (referred to as the 'best model') from among the candidate set. In other words, it can be used to identify the hypothesis that is best supported by observations. Two fundamentally different approaches are frequently used to address this in ecology and evolution (Box 3). One is to use a series of null hypothesis tests, such as LIKELIHOOD RATIO TESTS in phylogenetic analysis or F–tests in multiple regression analysis, to compare pairs of models from among the candidate set. However, this approach is typically restricted to nested models (i.e. the simpler model is a special case of the more complex model) and, in some cases, leads to suboptimal models that are dependent upon the hierarchical order in which models are compared [8]. Moreover, such tests cannot be used to quantify the relative support for the various models.

By contrast, model selection criteria can be used to rank competing models and to weigh the relative support for each one. These techniques utilize maximum likelihood scores as a measure of fit (more precisely, negative

Glossary

Akaike information criterion (AIC): an estimate of the expected Kullback–Leibler information [3] lost by using a model to approximate the process that generated observed data (full reality). AIC has two components: negative log-likelihood, which measures lack of model fit to the observed data, and a bias correction factor, which increases as a function of the number of model parameters.
Akaike weight: the relative likelihood of the model given the data. Akaike weights are normalized across the set of candidate models to sum to one, and are interpreted as probabilities. A model whose Akaike weight approaches 1 is unambiguously supported by the data, whereas models with approximately equal weights have a similar level of support in the data. Akaike weights provide a basis for model averaging (Box 4).
Least squares: a method of fitting a model to data by minimizing the squared differences between observed and predicted values.
Likelihood ratio test: a test frequently used to determine whether data support a fuller model over a reduced model (Box 3). The fuller model is accepted as best when the likelihood ratio (reduced model negative log-likelihood: full model negative log-likelihood) is sufficiently large that the difference is unlikely to have occurred by chance (i.e. $P < 0.05$).
Maximum likelihood: a method of fitting a model to data by maximizing an explicit likelihood function, which specifies the likelihood of the unknown parameters of the model given the model form and the data. Parameter values associated with the maximum of the likelihood function are termed the maximum likelihood estimates of that model.
Model averaging: a procedure that accounts for model selection uncertainty (defined below) in order to obtain robust estimates of model parameters ($\hat{\theta}$) or model predictions ($\hat{y}$) (Box 4). A weighted average of the model-specific estimates of $\hat{\theta}$ or $\hat{y}$ is calculated based on the Akaike weight [3] (or posterior probabilities if estimated using a Bayesian approach [48]) of each model. Where $\hat{\theta}$ does not appear in a model, the value of zero is entered.
Model selection bias: bias favoring models with parameters that are overestimated; such bias can be overcome during model averaging by entering the value 0 for parameters when they are not already included in the particular models to be averaged.
Model selection uncertainty: uncertainty about parameter estimates or model predictions that arises from having selected the model based on observations rather than actually knowing the best approximating model. Model selection uncertainty can be accounted for using model averaging.
Parametric bootstrap: a statistical technique in which new data are generated from Monte Carlo simulations of the fitted model. A measure of fit (typically the deviance) is then computed, both for the model fit to the observed data, and for the model fit to the simulated data. If the deviance of the model fit to the observed data falls within the core of the distribution of the deviance of model fit to the simulated data, then the model is said to fit the data adequately.
Parsimony: in statistics, a tradeoff between bias and variance. Too few parameters results in high bias in parameter estimators and an underfit model (relative to the best model) that fails to identify all factors of importance. Too many parameters results in high variance in parameter estimators and an overfit model that risks identifying spurious factors as important, and that cannot be generalized beyond the observed sample data.
Schwarz criterion (SC) (also known as the Bayesian information criterion) [10]: a model selection criterion designed to find the most probable model (from a Bayesian perspective) given the data (Box 3). Superficially similar to AIC_c, SC has two components: negative log-likelihood, which measures lack of fit, and a penalty term that varies as a function of sample size and the number of model parameters. SC is equivalent (under certain conditions) to the natural logarithm of the Bayes factor [48].

Box 3. Approaches to model selection

Once a set of candidate models is defined, they can be fit to observed data and compared to one another. Practitioners typically use one of three kinds of statistical approach to compare models: (i) maximizing fit; (ii) null hypothesis tests; and (iii) model selection criteria. Here, we highlight five frequently used techniques (Table I). Our list is not exhaustive (for additional examples, see [47–50]). Rather, we describe approaches most commonly used in ecology and evolutionary biology.

Maximizing fit

A naïve approach to model selection is to calculate a measure of fit, such as adjusted R^2, and select the model that maximizes that quantity. Maximizing fit, with no consideration of model complexity, always favors fuller (i.e. more parameter rich) models. However, it neglects the principle of PARSIMONY and, consequently, can result in imprecise parameter estimates and predictions, making it a poor technique for model selection. By contrast, tests or criteria that account for both fit and complexity are better suited for selecting a model.

Null hypothesis tests

The likelihood ratio test (LRT) is the most commonly used null hypothesis approach. LRT compare pairs of nested models. When the likelihood of the more complex model is significantly greater than that of the simpler model (as judged by a χ^2 statistic), the complex model is chosen, and vice versa. Selection of the more complex model indicates that the benefit of improved model fit outweighs the cost of added model complexity. LRT are often used hierarchically in a procedure analogous to forward selection in multiple regression, where the analyst starts with the simplest model and adds terms as LRTs indicate a significant improvement in fit. A drawback is that it requires several non-independent tests, thus inflating type I error. In addition, hierarchical LRTs sometimes select suboptimal models that are dependent upon the order in which models are compared, in which case dynamical LRTs can be employed [8]. However, no form of LRT can be used to quantify relative support among competing models.

Model selection criteria

Model selection criteria consider both fit and complexity, and enable multiple models to be compared simultaneously. The Akaike information criterion (AIC) estimates the Kullback–Leibler information lost by approximating full reality with the fitted model. Computation entails terms representing lack of fit and a bias correction factor related to model complexity. AIC has a second order derivative, AIC_c, which contains a bias correction term for small sample size, and should be used when the number of free parameters, p, exceeds $\sim n/40$ (where n is sample size). Schwarz criterion (SC; also referred to as a Bayesian information criterion, or BIC) [9] is structurally similar to AIC (Table I), but includes a penalty term dependent on sample size. Consequently, SC tends to favor simpler models, particularly as sample size increases [47]. Under certain conditions, model selection using SC and Bayes factor are equivalent, such that choosing the model with the smallest SC is equivalent to choosing the model with the greatest posterior probability [48]. Derivation of SC rests on several stringent assumptions that are seldom satisfied with empirical data, including that one true model exists, that this model is among the candidate set, and that the true model has an equal prior probability to each of the other models in the candidate set. Although SC superficially resembles AIC_c, it is not based in Kullback–Leibler information theory.

Which approach to use?

Which model selection approach is most appropriate? Techniques that maximize fit alone have clear limitations with regard to parsimony. Among approaches that consider fit and model complexity, many practitioners are moving from LRTs toward model selection criteria. For example, molecular systematists have traditionally used hierarchical LRTs to choose among competing models. However, this pattern could shift as researchers recognize the limitations of LRTs relative to the model selection criteria [4] (Box 5). Among model selection criteria, AIC is generally favored because it has its foundation in Kullback–Leibler information theory [3]. Yet, some prefer SC over AIC because the former selects simpler models [6]. An important advantage of using model selection criteria (e.g. AIC and SC) is that they can be used to make inferences from more than one model, something that cannot be done using the fit maximization or null hypothesis approaches.

Table I. Commonly used model selection methods

Model selection method	Calculation[a]	Elements	Refs
Adjusted R^2	$R^2_{adj} = 1 - \frac{RSS/n-p-1}{\sum(y_i-\bar{y})^2/n-1}$	Fit	[7]
Likelihood ratio test	$LRT = -2\{\ln[L(\hat{\theta}_p\|y] - \ln[L(\hat{\theta}_{p+q}\|y)]\} \sim \chi^2_q$	Fit and complexity	[7]
Akaike information criterion (AIC)	$AIC = -2\ln[L(\hat{\theta}_p\|y] + 2p$	Fit and complexity	[3]
Small sample unbiased AIC (AIC_c)	$AIC_c = -2\ln[L(\hat{\theta}_p\|y] + 2p\left(\frac{n}{n-p-1}\right)$	Fit and complexity (with bias correction term for small sample size)	[3]
Schwarz criterion	$SC = -2\ln[L(\hat{\theta}_p\|y] + p\cdot\ln(n)$	Fit, complexity, and sample size	[10]

[a]RSS, residual sum of squares for a linear model; n, sample size; p, count of free parameters (σ^2 must be included if it is estimated from the data); q, additional parameters of a fuller model; y: data; $L(\hat{\theta}|y)$: likelihood of the model parameters (more precisely, their maximum likelihood estimates, $\hat{\theta}_p$) given the data, y; for a model fitted by least squares with the usual assumptions, $\ln[L(\hat{\theta}_p|y)] = -n/2\ln(RSS/n)$, enabling computation of LRTs, AIC, AIC_c, and SC from standard regression output.

log-likelihood scores as a measure of lack of fit) and a term that, in effect, penalizes models for greater complexity. Two criteria commonly used in ecology and evolution are the AKAIKE INFORMATION CRITERION (AIC) [9] and the SCHWARZ CRITERION (SC; known also as the Bayesian information criterion, or BIC) [10]. The use of model selection criteria enable inference to be drawn from several models simultaneously, so that researchers can consider a 'best set' of similarly supported models.

Parameter estimation and model averaging

Often, the underlying motive for model selection is to estimate model parameters that are of particular biological interest (e.g. survival rate in mark–recapture studies, or transition:transversion ratios for phylogenetic studies), or to identify a model that can be used for prediction. When there is clear support for one model, maximum likelihood parameter estimates or predictions from that model can be used. However, there is sometimes nearly equivalent support in the observed data for multiple

Box 4. Multi-model inference

The model selection paradigm is moving beyond simply choosing a single, best model. Multi-model inference refers to a set of analysis techniques employed to enable formal inference from more than one model [3]. These techniques can be divided into two areas.

Generating a confidence set of models

How do we know which models are well supported by the data? A set of calculations based on Akaike information criterion (AIC) provides one way for making this determination. Once each model has been fit to the data and an AIC score has been computed, differences in these scores between each model and the best model are calculated (the 'best' model in the set has the minimum AIC score) (Eqn I)

$$\Delta_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min} \quad \text{[Eqn I]}$$

The likelihood of a model, g_i, given the data, y, is then calculated as Eqn II,

$$L(g_i|y) = \exp(-1/2\Delta_i) \quad \text{[Eqn II]}$$

In some cases, it is informative to contrast the likelihood of pairs of models, particularly that of the best model with each other model, using the evidence ratio (Eqn III),

$$ER = \frac{L(g_{best}|y)}{L(g_i|y)}. \quad \text{[Eqn III]}$$

Model likelihood values can also be normalized across all R models so that they sum to 1 (Eqn IV),

$$W_i = \frac{\exp(-1/2\Delta_i)}{\sum_{j=i}^{R} \exp(-1/2\Delta_j)} \quad \text{[Eqn IV]}$$

This value, referred to as the Akaike weight, provides a relative weight of evidence for each model. Akaike weights can be interpreted as the probability that model i is the best model for the observed data, given the candidate set of models. They are additive and can be summed to provide a confidence set of models, with a particular probability that the best approximating model is contained within the confidence set. They also provide a way to estimate the relative importance of a predictor variable (or a functional form that represents some biological process). This measure of relative importance can be calculated as the sum of the Akaike weights over all of the models in which the parameter (or functional form) of interest appears [3].

Model averaging

When the underlying goal of model selection is parameter estimation or prediction, and no single model is overwhelmingly supported by the data (i.e. $w_{best} < 0.9$), then model averaging can be used. This entails calculating a weighted average of parameter estimates, $\hat{\theta}$ (Eqn V),

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i \quad \text{[Eqn V]}$$

(where $\hat{\theta}_i$ is the estimate of $\hat{\theta}$ from the ith model) across all R models in the candidate set. The variance of these estimates can also be calculated (Eqn VI),

$$\hat{var}(\hat{\bar{\theta}}) = \sum_{i=1}^{R} w_i[\hat{var}(\hat{\theta}|g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2] \quad \text{[Eqn VI]}$$

(where $\hat{var}(\hat{\theta}|g_i)$ is the estimate of the variance of θ from the ith model). This variance estimator can be used to assess the precision of the estimate over the set of models considered, thereby providing a way to generate a confidence interval on the parameter estimate that accounts for model selection uncertainty. Predicted values of the response variable can be averaged over the models in the candidate set in an analogous way [3].

models [i.e. Akaike information criterion (AIC) values are nearly equal], making it problematic to choose one model over another. MODEL AVERAGING provides a way to address this problem (Box 4). Parameter estimates or predictions obtained by model averaging are robust in the sense that they reduce MODEL SELECTION BIAS and account for MODEL SELECTION UNCERTAINTY.

Inference from model selection

Ultimately, model selection is a tool for making inference about unobserved processes based on observed patterns. Data that clearly support one model over several others lend strong support to the corresponding hypothesis (among those considered); that is, we can infer the process that is most likely to have operated in generating the observed data. However, some inferences, such as determining the relative importance of predictor variables, can be made only by examining the entire set of candidate models (Box 4).

Where model selection is being used

Model selection is well established as a basic tool in select biological disciplines. In particular, it is a prerequisite for most mark–recapture studies and for most phylogenetic studies (Box 5). Model selection is now beginning to be implemented more broadly to address a variety of additional questions in ecology and evolution (Table 1). Here, we highlight some areas where such an approach has proved useful.

Ecology

Mark–recapture analyses are used widely to estimate population abundance and survival probabilities [11,12]. A fundamental challenge is to separate the probability that a marked individual has died from the probability that it was not recaptured in spite of having survived. Wildlife biologists address this problem by generating a set of competing models that depict different ways in which survival and encounter probabilities could vary as a function of time, the environment, or individual traits (e.g. sex or size) (Box 5). The favored model (or set of models) is then used to estimate parameters of interest, or to infer the biological processes governing survival or abundance. This approach has been used to estimate vital rates for management and conservation [13,14], and to infer how factors, such as individual physiological status, or environmental conditions, affect vital rates [15,16]. Community ecologists [17] and paleontologists [18] have even adopted this mark–recapture model selection framework to estimate species richness and species turnover rates.

There is also a rich tradition of using models to explore population dynamics [6]. Ecologists have proposed many competing hypotheses to explain patterns of population fluctuation over time. An increasing number of studies have fit models depicting competing hypotheses to observed time series data; applications include detecting chaotic dynamics in natural populations [19], inferring the mechanism underlying population cycles [20,21], and separating the influence of density-dependent and

Box 5. Parallel development of model selection in wildlife biology and molecular systematics

Although the initial statistical machinery and philosophical underpinnings of model selection have been available for 30 years [9], ecologists and evolutionary biologists have only recently expanded and incorporated this tool. Wildlife biologists and molecular systematists have been at the forefront of bringing model selection to ecology and evolution, yet the approach has been applied almost independently in these two fields. Still, there are striking similarities and interesting differences in how model selection is currently used (Table I).

Wildlife biology

Fifteen years ago, a group of wildlife biologists grappling with the problem of how to compare non-nested models began using the Akaike information criterion (AIC) as a basis for model selection [11]. Consequently, AIC_c (or its variant $QAIC_c$ used for overdispersed count data) is now standard in mark–recapture analysis [45]. Goodness-of-fit testing and model averaging also are commonly used in mark–recapture studies. Most recently, the trend is toward using multiple models to estimate parameters of interest and to infer biological processes. Hence, hierarchical likelihood ratio tests (LRT) are seldom employed.

Molecular systematics

Molecular systematists found a need for model selection because different models of DNA sequence evolution sometimes result in the construction of different trees [51]. Hence, over the past ten years, a view has evolved among many systematists that it is necessary to identify one best-fitting model from a nested set of candidate models, and then use this chosen model to generate the phylogeny [46]. Goodness-of-fit testing is rare in systematics, and hierarchical LRTs remain common. However, interest in AIC, and its broader utility in molecular systematics, appears to be increasing [4].

Integrating across fields

Recent interactions between wildlife biology and molecular systematics in the use of model selection are leading to exciting new developments. For example, a primary focus of mark–recapture studies is to estimate survival rates, where model averaging is used to yield more robust estimates of model parameters. Molecular systematists frequently use estimates of model parameters in phylogeny reconstruction, but have traditionally relied on maximum likelihood estimates from a single best model. However, using model averaging to obtain more robust parameter estimates provides a new option in phylogeny reconstruction [4]. Similarly, Akaike weights could be used to determine the relative support for conflicting topologies generated under different models of molecular evolution, and might provide a basis for combining discordant trees [4]. Hence, the integration of model selection techniques across disciplines, particularly multi-model inference (Box 4), promises to bring together several previously distinct fields.

Table I. Comparison of model selection implementation in mark-recapture research and molecular systematics[a]

	Mark–recapture studies	Molecular systematics
Objective	To estimate parameters (survival rates, recapture rates, and transition rates) based on recovery of marked individuals	To identify a model of molecular evolution and model parameter estimates that can be used in phylogenetic reconstruction
Model types	Multinomial probability models	Multinomial probability models
Set of candidate models	Parameter families [10]: S, survival probability p, detection probability ψ, transition probability (multi-strata models) Model variations: Parameter constant, $\theta_\bullet$ Parameter varying freely over time, θ_t Parameter differing among groups, θ_g Parameter differing by patch, θ_r Linear trend in parameter value, $\theta = f(t)$ Parameter a function of a covariate, $\theta = f(x)$	Parameter families [46]: τ, phylogenetic tree, including branch lengths π, nucleotide base frequencies I, proportion of invariable nucleotide sites in a set of aligned DNA sequences Γ, substitution rate heterogeneity among nucleotide sites (gamma distribution with four discrete categories) ϕ, substitution rate variation among nucleotides (6 classes of transitions and transversions)
Goodness of fit test	Commonly used; applied to the most complex model before the model selection step	Very rare; when used, applied to the best model after the model selection step [52]
Model fitting algorithm	Maximum likelihood	Maximum likelihood
Model selection criterion	Predominantly AIC_c or $QAIC_c$; LRT seldom used	Predominantly hierarchical LRT; AIC seldom used
Use of model averaging	Uncommon, but available and sometimes used [3]	Recently introduced, but still rarely used [4]
Software commonly used	*MARK* [45]	*MODELTEST* [46]

[a]Abbreviations: AIC, Akaike information criterion; LRT, likelihood ratio test; QAIC, variant of AIC for overdispersed count data.

environmental factors [22]. However, in spite of a heavy reliance on AIC for model selection in statistical time series analysis, only recently have population ecologists applied model selection to quantify support for competing explanations [23], an approach that appears to be promising as a way to infer mechanisms that control natural fluctuations in population size.

Evolution

Model selection now underpins most phylogenetic reconstruction. All methods of phylogenetic inference are based on hypotheses about how biological characters change through time [24]. When phylogenies are reconstructed from DNA data, these hypotheses can be expressed as competing models of nucleotide substitution [25] (Box 5). In molecular phylogenetics, it is now common to consider multiple models of molecular evolution before selecting a single best model to be used in maximum likelihood or Bayesian phylogenetic reconstruction [8,26,27]. Recent advances in model-based morphological phylogenetics [28,29] suggest that model selection can also be used to address a variety of new questions relating to the

Table 1. Increasing use of model selection in ecology and evolution

Discipline	Problem	Refs
Ecology		
Natural history	Identifying foraging strategies of species (generalist versus specialist)	[53]
Population ecology and management	Isolating endogenous and exogenous mechanisms of regulation	[23,54]
	Detecting spatial heterogeneity in population regulation	[55]
	Relating survival rates to physiological and environmental factors (mark–recapture data)	[13–16]
	Correlating vital rates with covariates (monitoring data)	[56]
	Modeling herbivore functional response	[57]
Behavioral ecology	Discerning how animals allocate risk in response to predation	[58]
	Modeling dispersal	[59]
Community ecology	Modeling effects of fire on community organization	[60]
Landscape ecology	Predicting how vertebrate populations respond to habitat loss and fragmentation	[61]
Ecosystem science	Deciphering trophic relationships	[40]
Evolution		
Molecular evolution	Understanding the process of nucleotide/protein evolution	[62,63]
Molecular systematics	Choosing a model of molecular evolution for phylogenetic reconstruction	[4,64,65]
Life history evolution	Identifying selective agents associated with phenotypes	[30,31]
Adaptive radiation	Estimating historical diversification rates of lineages	[66]
Genetic mapping	Identifying the genetic architecture of phenotypes	[67]
Population genetics	Examining patterns of gene flow	[68]
Historical demography	Using genetic markers to infer past population dynamics	[69]

rate and patterns of morphological character evolution over time.

A more recent application of model selection in evolutionary biology is to identify selective pressures that shape adaptations in the wild. Given the complexity of natural systems, there are often several ecological factors and a variety of mechanisms that could explain evolutionary change. Fitting competing models to observed data can represent these alternative explanations. For example, model selection has been used recently to explore probable causes of life-history diversification in natural systems, including body size at emergence and timing of emergence in desert stream caddisflies [30] and size at maturity, number and size of offspring, and reproductive investment in tropical live-bearing fish [31].

When should model selection be used?

Model selection is well suited for making inferences from observational data, especially when data are collected from complex systems or when inferring historical scenarios where several different competing hypotheses can be put forward. Not surprisingly, such conditions are typical of many research problems in ecology and evolution, particularly when experimental manipulation is not possible. Unfortunately, null hypothesis testing remains the dominant mode of inference in ecology and evolution [2], even for studies that are best suited to the model selection approach. We illustrate this with two examples.

Statistical phylogeography

A goal of phylogeography is to uncover the geographical and demographic histories of populations [32,33]. Given that it is impossible to test population histories experimentally, inferences must be made using contemporary genetic data: typically observations of the spatial distribution of genetic variation among extant populations, combined with gene trees. Recent work has highlighted the advantages of statistically testing multiple historical scenarios [34–36]. Yet, the statistical framework has been limited to null hypothesis tests. Such approaches yield a single population history, but fail to provide insights into estimate error and do not consider the relative support for alternative scenarios. Some statistical phylogeographers, aware of this shortcoming, have recently called for an approach that promotes the generation of explicit models of population histories, whilst providing the tools to evaluate the fit of these models to observed data [36]. Model selection could provide a statistical framework to help fill this void.

Ecosystem science

A focal problem in ecosystem science is unraveling complex trophic relationships among taxa. This issue has been addressed at both the theoretical [37] and empirical [38] level using models of food chains and food webs. The current state-of-the-art in ecosystem modeling is to advance a simple hypothesis, to acquire a few observational data sufficient to test the simple hypothesis, and to use these results to show where the assumptions of the simple model failed, thus leading to a refined hypothesis and further testing [39]. Model selection offers a framework through which empirical support for a set of food-web models can be weighed simultaneously. The utility of this approach was demonstrated in a study of subterranean interactions among plants, root-feeding caterpillars, and nematode parasitoids of the caterpillars. Model selection revealed that nematodes provided the shrubs an appreciable degree of protection from caterpillars, a result whose ecological interpretability would not have been attained using the conventional logistic regression approach [40]. Hence, adopting model selection appears to hold great promise for increasing our understanding of trophic interactions, and should have similar utility in other systems that are too complex for experimental manipulation.

Caveats and future direction

As the use of model selection becomes more widespread, it is important to be aware of potential pitfalls and

opportunities for future development. We offer three ideas. First, inferences derived from model selection ultimately depend on the models included in the candidate set. Hence, failure to include models that might best approximate the underlying biological process [41–43], or spurious inclusion of meaningless models, could each lead to misguided inference. Therefore, researchers must think critically about alternative biological hypotheses before data are collected and analyzed. Second, if a model is to carry biological meaning, rather than mere statistical significance, then its predictions and parameter estimates must be biologically plausible. Thus, models that fail to predict known patterns, or those that generate implausible estimates should be viewed as untenable [30]. In other words, it is logically inconsistent to accept empirical support for a model and its associated hypothesis (e.g. using AKAIKE WEIGHTS) whilst discarding its parameter estimates and predictions. Finally, biologists must decide when it is most appropriate to use model selection, and when it is most appropriate to use designed experiments and inferences based on significance tests. Certain phenomena, such as the evolutionary diversification of a lineage over tens of thousands of years, are clearly beyond the reach of controlled experiments; inference based on model selection is the only option in such cases. Other phenomena, such as population cycling, can be studied using observational time series data [21] or by manipulative experimentation [44], sometimes creating conflict as to which approach is most fruitful. Given recent advances in model-based inference, the complementary utility of these two approaches warrants further attention.

The potential for model selection to be applied to many more problems in ecology and evolutionary biology is exciting. The model selection paradigm makes it clear when the data show equivocal support for more than one hypothesis. Practitioners accustomed to statistical hypothesis tests that generate either 'significant' or 'nonsignificant' results might be frustrated that a single answer does not always emerge. Yet, this ability to weight evidence for competing hypotheses is precisely the strength of model selection. Moreover, identifying levels of support for competing hypotheses appears to be only a start for how this tool might ultimately be employed. Advances in multimodel inference promise to broaden the usefulness of the model selection paradigm. As model selection matures, we anticipate that it will continue to spread in ecology and evolution, expanding the set of statistical tools available to researchers.

Acknowledgements

Our thanks go to David Anderson, Nick Gotelli, David Lytle, Kevin Omland, and David Posada for helpful comments about the article, and to David Anderson for providing equation VI. Funding from a National Research Council Research Associateship Award to J.B.J. and the Vermont Cooperative Fish and Wildlife Research Unit to K.S.O. generously supported the authors during the writing of this review.

References

1 Hilborn, R. and Mangel, M. (1997) *The Ecological Detective: Confronting Models With Data*, Princeton University Press
2 Anderson, D.R. *et al.* (2000) Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Manage.* 64, 912–923
3 Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer
4 Posada, D. (2003) Unit 6.5: Using MODELTEST and PAUP* to select a model of nucleotide substitution. In *Current Protocols in Bioinformatics* (Vol. 1) (Baxevanis, A.D. *et al.*, eds), pp. 6.5.1–6.5.28, John Wiley & Sons
5 Chamberlain, T.C. (1890) The method of multiple working hypotheses. *Science* 15, 92–96
6 Turchin, P. (2003) *Complex Population Dynamics: A Theoretical/Empirical Synthesis*, Princeton University Press
7 Sokal, R.R. and Rohlf, F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*, W.H. Freeman & Co
8 Posada, D. and Crandall, K.A. (2001) Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50, 580–601
9 Akaike, H. (1973) Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Petrov, B.N. and Csaki, F., eds), pp. 267–281, Akademiai Kiado
10 Schwarz, G. (1978) Estimating the dimensions of a model. *Ann. Stat.* 6, 461–464
11 Lebreton, J.D. *et al.* (1992) Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecol. Monogr.* 62, 67–118
12 Schwarz, C.J. and Seber, G.A.F. (1999) Estimating animal abundance: review III. *Stat. Sci.* 14, 427–456
13 Schreiber, E.A. *et al.* (2001) Effects of a chemical weapons incineration plant on red-tailed tropicbirds. *J. Wildl. Manage.* 65, 685–695
14 Sillett, T.S. and Holmes, R.T. (2002) Variation in survivorship of a migratory songbird throughout its annual cycle. *J. Anim. Ecol.* 71, 296–308
15 Jorgenson, J.T. *et al.* (1997) Effects of age, sex, disease, and density on survival of bighorn sheep. *Ecology* 78, 1019–1032
16 Esler, D. *et al.* (2000) Winter survival of adult female harlequin ducks in relation to history of contamination by the *Exxon Valdez* oil spill. *J. Wildl. Manage.* 64, 839–847
17 Boulinier, T. *et al.* (1998) Estimating species richness: the importance of heterogeneity in species detectability. *Ecology* 79, 1018–1028
18 Connolly, S.R. and Miller, A.I. (2001) Joint estimation of sampling and turnover rates from fossil databases: capture-mark-recapture methods revisited. *Paleobiology* 27, 751–767
19 Ellner, S. and Turchin, P. (1995) Chaos in a noisy world: new methods and evidence from time-series analysis. *Am. Nat.* 145, 343–375
20 Kendall, B.E. *et al.* (1999) Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology* 80, 1789–1805
21 Turchin, P. and Hanski, I. (2001) Contrasting alternative hypotheses about rodent cycles by translating them into parameterized models. *Ecol. Lett.* 4, 267–276
22 Dennis, B. and Otten, M.R.M. (2000) Joint effects of density dependence and rainfall on abundance of San Joaquin kit fox. *J. Wildl. Manage.* 64, 388–400
23 White, G.C. and Lubow, B.C. (2002) Fitting population models to multiple sources of observed data. *J. Wildl. Manage.* 66, 300–309
24 Felsenstein, J. (2003) *Inferring Phylogenies*, Sinauer Associates
25 Felsenstein, J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22, 521–565
26 Huelsenbeck, J.P. and Crandall, K.A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28, 437–466
27 Huelsenbeck, J.P. and Rannala, R. (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276, 227–232
28 Lewis, P.O. (2001) Phylogenetic systematics turns over a new leaf. *Trends Ecol. Evol.* 16, 30–37
29 Lewis, P.O. (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50, 913–925
30 Lytle, D.A. (2002) Flash floods and aquatic insect life-history evolution: evaluation of multiple models. *Ecology* 83, 370–385
31 Johnson, J.B. (2002) Divergent life histories among populations of the fish *Brachyrhaphis rhabdophora*: detecting putative agents of selection by candidate model analysis. *Oikos* 96, 82–91
32 Avise, J. (2000) *Phylogeography: The History and Formation of Species*, Harvard University Press
33 Hare, M.P. (2001) Prospects for nuclear gene phylogeography. *Trends Ecol. Evol.* 16, 700–706

34 Templeton, A.R. (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol. Ecol.* 7, 381–397
35 Wakeley, J. and Hey, J. (1998) Testing speciation models with DNA sequence data. In *Molecular Approaches to Ecology* (DeSalle, R. and Schierwater, B., eds), pp. 157–175, BirkVerlag-Verlag
36 Knowles, L.L. and Maddison, W.P. (2002) Statistical phylogeography. *Mol. Ecol.* 11, 2623–2635
37 Williams, R.J. and Martinez, N.D. (2000) Simple rules yield complex food webs. *Nature* 404, 180–183
38 Carpenter, S.R. *et al.* (1999) Management of eutrophication for lakes subject to potentially irreversible change. *Ecol. Appl.* 9, 751–771
39 Power, M.E. (2001) Field biology, food web models, and management: challenges of context and scale. *Oikos* 94, 118–129
40 Strong, D.R. *et al.* (1999) Model selection for a subterranean trophic cascade: root-feeding caterpillars and entomopathogenic nematodes. *Ecology* 80, 2750–2761
41 Dreitz, V.J. *et al.* (2001) Spatial and temporal variability in nest success of snail kites in Florida: a meta-analysis. *Condor* 103, 502–509
42 Beissinger, S.R. and Snyder, N.F.R. (2002) Water levels affect nest success of the snail kite in Florida: AIC and the omission of relevant candidate models. *Condor* 104, 208–215
43 Dreitz, V.J. *et al.* (2002) Snail kite nest success and water levels: a reply to Beissinger and Snyder. *Condor* 104, 216–221
44 Krebs, C.J. *et al.* (1995) Impact of food and predation on the snowshoe hare cycle. *Science* 269, 1112–1115
45 White, G.C. and Burnham, K.P. (1999) Program MARK: survival estimation from populations of marked animals. *Bird Stud.* 46, S120–S139
46 Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (Hillis, D. *et al.*, eds), pp. 407–514, Sinauer Associates
47 Zucchini, W. (2000) An introduction to model selection. *J. Math. Psychol.* 44, 41–61
48 Wasserman, L. (2000) Bayesian model selection and model averaging. *J. Math. Psychol.* 44, 92–107
49 Suchard, M.A. *et al.* (2001) Bayesian selection of continuous-time Markov Chain evolutionary models. *Mol. Biol. Evol.* 18, 1001–1013
50 Huelsenbeck, J.P. *et al.* (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310–2314
51 Sullivan, J. and Swofford, D.L. (1997) Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J. Mamm. Evol.* 4, 77–86
52 Sullivan, J. *et al.* (2000) Comparative phylogeography of mesoamerican highland rodents: concerted versus independent response to past climatic fluctuations. *Am. Nat.* 155, 755–768
53 Luh, H.K. and Croft, B.A. (1999) Classification of generalist or specialist life styles of predaceous phytoseiid mites using a computer genetic algorithm, information theory, and life history traits. *Environ. Entomol.* 28, 915–923
54 Erb, J. *et al.* (2001) Spatial variation in mink and muskrat interactions in Canada. *Oikos* 93, 365–375
55 LaMontagne, J.M. *et al.* (2002) Spatial patterns of population regulation in sage grouse (*Centrocercus* spp.) population viability analysis. *J. Anim. Ecol.* 71, 672–682
56 Pease, C.M. and Mattson, D.J. (1999) Demography of the Yellowstone grizzly bears. *Ecology* 80, 957–975
57 Hobbs, N.T. *et al.* (2003) Herbivore functional response in heterogeneous environments: a contest among models. *Ecology* 84, 666–681
58 Van Buskirk, J. *et al.* (2002) A test of the risk allocation hypothesis: tadpole responses to temporal change in predation risk. *Behav. Ecol.* 13, 526–530
59 Zabel, R.W. (2002) Using 'travel time' data to characterize the behavior of migrating animals. *Am. Nat.* 159, 372–387
60 Beckage, B. and Stout, I.J. (2000) Effects of repeated burning on species richness in a Florida pine savanna: a test of the intermediate disturbance hypothesis. *J. Veg. Sci.* 11, 113–122
61 Swihart, R.K. *et al.* (2003) Responses of 'resistant' vertebrates to habitat loss and fragmentation: the importance of niche breadth and range boundaries. *Div. Distrib.* 9, 1–18
62 Yang, Z. *et al.* (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431–449
63 Posada, D. and Crandall, K.A. (2001) Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.* 18, 897–906
64 Jordan, S. *et al.* (2003) Molecular systematics and adaptive radiation of Hawaii's endemic damselfly genus *Megalagrion* (Odonata: Coenagrionidae). *Syst. Biol.* 52, 89–109
65 Buckley, T.R. *et al.* (2002) Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. *Syst. Biol.* 51, 4–18
66 Paradis, E. (1998) Detecting shifts in diversification rates without fossils. *Am. Nat.* 152, 176–187
67 Sillanpää, M.J. and Corander, J. (2002) Model choice in gene mapping: what and why. *Trends Genet.* 18, 301–307
68 Roach, J.L. *et al.* (2001) Genetic structure of a metapopulation of black-tailed prairie dogs. *J. Mammal.* 82, 946–959
69 Strimmer, K. and Pybus, O.G. (2001) Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol. Biol. Evol.* 18, 2298–2305